Journal of Second Language Teaching and Research. Volume 7

Comparing lexical development at two distinct IELTS bands within an international foundation programme

Andrew Drummond, *King's College London*

Abstract

This study measures lexical development in the writing of two groups of non-native speaking students on an international foundation programme at a UK University. The higher-level group entered the programme with IELTS 7.0 in writing and the lower-level group 5.5 in writing. Laufer and Nation's (1995) Lexical Frequency Profile has been used, along with Antwordprofiler (Anthony, 2014), to calculate what proportions of common and less common vocabulary were present in their writing at the beginning and end of the academic year. These proportions were then compared with a benchmark taken from a corpus of 30 essays of accomplished students' writing. The results show that the higher-level group moved firmly into the range of the native-speaker benchmark, but the lower-level group made more limited progress. Other measurements based on lexical variation give a different picture of lexical development in the lower group, indicating that lexical knowledge should be treated as a multi-dimensional construct. Implications for EAP courses are discussed.

Keywords: *lexical frequency profile; IELTS writing scores; foundation programme; lexical sophistication; EAP; students' writing*

Introduction

Many UK Universities run foundation programmes for international students who wish to study in the UK but require an additional year of schooling and/or language skills development before commencing their undergraduate courses. Students entering foundation programmes are often grouped according to the scores they achieve on an IELTS test. An IELTS score is essentially a 'manageable proxy [measure] of academic readiness for mainstream university study' (Singh & Doherty, 2004:10). Foundation programmes may attract students with diverse levels of language development, requiring the same foundation course to cater for students with IELTS scores of 5.0 overall at the lower range and 8.0/8.5 overall at the upper range. Providing programmes that are adequately structured to meet such diverse needs is a challenge for educators.

One such problem is deciding how much language input the higher-level students require. IELTS guidelines (IELTS, 2015:25) state that a score of 7.0, 'will probably meet the language requirements of most university courses. In light of this, EAP practitioners (and students) may feel that those with IELTS scores of 7.0 or above are already so linguistically developed that no further development of written language or spoken language would be necessary before commencing undergraduate study. Yet the literature on the readiness of IELTS 7.0 students to commence undergraduate study is by no means exhaustive. We do not know, for example, how the lexical knowledge of IELTS 7.0 students compares with native-speaker undergraduates. It would also be advantageous to know how students with lower IELTS scores (5.5, for example) compare in terms of their productive lexical capabilities. It would seem advantageous for practitioners involved with international foundation programmes to have clearer data on the academic writing of these diverse types of students.

The production of academic writing is complex, with many linguistic and study skills resources being required. One such resource is the possession of a sufficient productive lexicon. Nation and Laufer (1995:307) state that: 'a well-used rich vocabulary is likely to have a positive effect on the reader' and this was also found to be the case by Yu (2010). Moreover, fluency with the specialist terminology of an academic discipline is key to participating in the discourse of that particular community (Corson, 1997). Academic text, being 'lexically dense' (Halliday, Matthiessen & Matthiessen, 1985:61), is dependent on an adequate lexicon for its construction. The learning of academic and subject-specific vocabulary forms an important part of foundation courses for international students. Sufficient lexical knowledge, then, is a key to writing to an acceptable institutional standard.

Measuring and tracking lexical development in foundation students forms the focus of this study. In particular, data on the lexical development of two groups of students over the course of a one-year foundation programme were collected. The two groups were those with IELTS 5.5 in writing at the start of the course (IELTS 5.5 group) and those with IELTS 7.0 in writing at the start of the course (IELTS 7.0) group. Lexical development data were tracked with computer software. These data were compared with a benchmark of the 'productive lexical level' evidenced in a corpus of accomplished students' writing native speakers (see below for how this benchmark was established). The term 'productive lexical level' is defined in this paper as the size of a student's productive lexicon, to the extent that it is measurable in their written texts.

The investigation reveals how close the IELTS 5.5 group and the IELTS 7.0 group are to native-speaker levels of productive lexical knowledge at the beginning and end of the course. Native speakers, in this study, refers to students who have self-reported English as their first language when submitting their essays to the corpora represented in this study. This data could be used to provide information on how learning programmes ought to be structured. In particular, the validity of any institutional assumption that higher level students do not need additional language input is investigated.

Literature Review

Measuring productive lexical knowledge

The Lexical Frequency Profile (LFP) was suggested by Nation and Laufer (1995) as a means of assessing productive lexical level. An LFP essentially compares the proportion of frequent words to infrequent words in a particular piece of writing. The assumption is that a richer lexicon will be evident in a 'larger proportion of infrequent words … in a text' (Laufer, 2012:3). This correlation in the size of lexicon and proportion of infrequent words was established by Nation and Laufer (1995) who corroborated students' scores on a paper-based vocabulary text with the LFP evident in their written work. Nation and Laufer (ibid) initially calculated the LFP with computer software called VocabProfile, utilizing West's (1953) General Service List (GSL) and Xue and Nation's (1984) University Word List (UWL). This created data on the proportion of words in text appearing in each of these word lists; the GSL represented the most common 1000 word families and the second most common 1000 word families in general English and UWL was an early academic word list. In much subsequent LFP-based research, the UWL was largely replaced by Coxhead's (2000) Academic Word List (AWL).

The software used for the present study is AntWordProfiler (Anthony, 2014). It provides rich data on the lexical composition of texts including counts of word 'tokens', 'types' and 'families' at the levels of frequency commensurate with the set of word lists used. The measure of word *tokens* in a text is a count of the total number of running words from a particular list. The measure of word *types*, however, refers to the number of unique words present at each strata of frequency (Loewen & Plonsky, 2015). The following example contains nine *tokens* but only eight *types* due to the repetition of the word 'of':

'This sentence contains a number **of** types **of** word.'

In contrast to word types and word tokens, a measurement of word *families* in a text is determined by counting words sharing the same base form. As Bauer & Nation (1993:253) state: 'a word family consists of a base word and all its derived and inflected forms'. The following made-up sentence, then, contains six *tokens* but only four word *families*.

'I **wrongly wronged** the **wrong** man.'

Since the software can count types, tokens and word families in a text, the researcher needs to decide which of these measures to use. Counting only word *tokens* in a text is inadequate since such a high proportion of infrequent lexis may be present as a result of the writer repeating the same limited number of 'infrequent' words several times. For example, the word 'lexis', which is relatively infrequent in the language at large, appears 12 times in this paper.

Counting word *types*, in contrast, provides data on the proportion of unique forms of a word in the texts without assuming that if one form is productively known, the whole family is known. In fact, Schmitt & Zimmerman (2002) questioned whether students' skills in morphological manipulation of the base form of a word should be assumed. Although Nation and Laufer (1995) focused on word families in their original study, this study will focus on measures of word types. Each measure has different merits but for this study, which contains higher and lower level writers, a measure of word types rather than word families will not make assumptions about lower level learners who use one or two derivations of a base form.

LFP has demonstrable advantages over other methods which have been used to measure lexical level, such as lexical originality, lexical density, and lexical sophistication (Nation & Laufer, 1995:308-11). Lexical originality is a way of measuring the number of unique words in a text, relative to group norms, so it is essentially a relativistic measure and cannot be used effectively to compare the lexical level of two cohorts with varying group norms. Similarly, lexical sophistication, which measures the proportion of advanced words in a text relative to the total number, is not effective unless there is a broad enough consensus on which words ought to be considered advanced and which basic (Nation and Laufer, 1995:308-11). Lexical density, which is a measure of the proportion of content words in a text relative to function words, is also considered an inadequate measure by Nation and Laufer (1995:309) since, 'it does not

necessarily measure lexis [but] depends on the syntactic and cohesive properties of the composition.' In other words, a text could be lexically packed with items from the most frequent 1000 word families and be an example of syntactic complexity rather than lexical complexity.

Another lexical measure that is considered limited is Type-Token Ratio (TTR). TTR is a measure of lexical diversity achieved by dividing the number of unique words in a text by the total number of running words. While TTR is known to be sensitive to text length (Loewen & Plonsky,2015), the process of standardising text length within a corpus has been shown to provide reliable results (Treffers-Daller, Parslow & Williams, 2016). Schmidt (2010) criticizes type-token ratios for failing to differentiate between levels of lexical frequency i.e. a text may score very positively using this measure by using lexis entirely from the 1000 most common words in the language.

Whilst LFP yields more reliable and comprehensive data than older measures, it also has limitations. For example, it provides no data on collocation and, as such, omits a key dimension from its analysis. In fact, any lexical item composed of more than one word, such as a lexical bundle (e.g. 'in order to be') or fixed prepositional phrase (e.g. 'in the main'), is not recognised by AntWordProfiler or Range (Heatley et al., 2002). Collocation accuracy has been shown to influence a reader's perception of the quality of a text (Crossley, Salsbury, & Mcnamara, 2014) and it is possible that a text showing a high proportion of infrequent words based on its LFP could contain collocation inaccuracies. Furthermore, where a frequent word has been used with a more nuanced meaning, such as 'house' used as a verb, it will feature among 1k word family data on the basis of its more common nominal usage. In this sense, LFP provides data on a particular dimension of lexical knowledge without providing data at the level of collocation and multi-word units or polysemy. Nonetheless, developing productive knowledge of lexical items consisting of only one word remains an important part of a developing lexicon and there is some evidence of links between ability with collocations and overall vocabulary size (Brown, 2012).

Related studies

As mentioned above, the AWL (Coxhead, 2000) and the GSL (West, 1953) have been widely used as word lists to establish LFPs. For example. Gregori-Signes & Clavel-Arroitia (2015) utilised these wordlists in their study of lexical richness in Spanish-speaking undergraduates' English in Valencia. They found that a reliable LFP calculation was obtainable across two distinct texts. Turlik (2013) demonstrated a significant increase in the proportion of AWL items

present in students' writing over the course of a foundation programme. Iwashita (2005) observed that higher level students produced a greater proportion of advanced vocabulary in speaking tasks.

Other studies, while not measuring LFP, have generated knowledge related to the present study. For example, Mazgutova & Kormos (2015) investigated syntactic and lexical development of two cohorts differentiated by IELTS score. They assessed students' essay writing at the beginning and end of a short intensive pre-sessional EAP course and found lexical development had occurred in this period. The present study differs from their investigation in the respect that the relative progress of two groups towards a native-speaker benchmark over a longer period is to be ascertained. Cooper (2013) compared the sophistication of lexical bundles present in an IELTS writing task with later pieces of writing from first year undergraduates. Her results (ibid) questioned the validity of the assumption that IELTS scores are reliable indicators of academic readiness. Similarly, Drummond (2018) showed a wide range of receptive vocabulary knowledge present within each band of the IELTS scale, with a notable number of IELTS 7.0 (overall score) students exhibiting markedly low levels of receptive vocabulary knowledge. In spite of these studies, there are no published articles to my knowledge measuring the development of LFP at two distinct levels (based on high and low IELTS writing scores at entry) of an international foundation programme.

Research Questions

These are the research questions explored in the following study:

1. What could be considered a benchmark of productive lexical usage as evidenced in a corpus of accomplished students' writing?
2. To what extent did lexical development occur in the IELTS 5.5 and IELTS 7.0 non-native speakers' groups over the course of the foundation year?
3. To what extent did the LFP of each non-native speaker group develop towards the LFP of successful native-speaker writing identified in research question 1?

Method

Word lists used in this study

Whilst the GSL and the AWL, as used in Laufer and Nation's seminal study (1995), have been very influential, they have been subject to criticism. The GSL, due to its age, contains lexical

items of nautical, agricultural and religious relevance that were current and frequent within West's corpora but are not similarly relevant today (Browne, 2014). In addition, it has been noted that the AWL contains a disproportionately high number of commerce and law-related items and a proportion of Coxhead's (2000) corpus consisted of texts from the Brown and LOB corpora which, as Hyland & Tse (2007) note, were considered dated even when they were writing. In response, Browne, Culligan, & Phillips (2013) have formulated a New General Service List (NGSL) and a New Academic Word List (NAWL), intended to improve on these noted weaknesses. The NGSL is based on a very large sample of 273 million words taken from the Cambridge English Corpus. This is a far larger and more modern language sample than the original 2.5 million words used to construct the GSL, 'reflect[ing] modern usage patterns' (Stoeckela & Bennett, 2015:2). The NGSL and NAWL, then, have been used for the present study to circumvent the shortcomings of the aging GSL. Word lists derived from this more modern usage should help to establish a more accurate picture of a student's productive lexicon and not unfairly assess them on the basis of language in the GSL which is no longer used.

The following table shows the NGSL and NAWL word lists used with AntWordProfiler in this study to calculate the proportion of frequent and infrequent words in the students' writing:

Table 1. *Wordlists used in the present study*

Sublist	Frequency level	Referred to as:
NGSL1	1-1000 most common words	1k
NGSL2	1001-2000	2k
NGSL3	2000-2818	n/a
NAWL	963	NAWL

Some studies into students' LFPs (Laufer, 1995; Lemmouh, 2008) have distinguished between the proportion of words among the 2000 most common in the language and the proportion of words beyond this threshold level of 2000. Laufer (1995) termed these less frequent words 'Beyond 2000' (B2000). The same distinction and terminology is used in this study, along with 'F2000' for the proportion of word types among the 2000 most frequent in English.

Assembling and analysing the benchmark corpus

This study aims to show how close two distinct groups of L2 students are to native-speaker level vocabulary use at the beginning and end of a foundation programme. In order to do this, a benchmark must first be established to measure native-speaker lexical use in writing. Data for this benchmark comes from a corpus of 30 essays: 15 from The Michigan Corpus of Upper-Level Student Papers (MICUSP) and 15 from the British Academic Written English (BAWE) corpus. MICUSP is a collection of A-graded student papers made available online for research purposes (Romer and Swales, 2010). The essays submitted to the BAWE corpus gained either a merit or a distinction. The advantage of taking essays from more than one institution is that it lessens the potential for idiosyncratic institutional practices encouraging either an especially lexically rich or lexically sparse form of writing and, therefore, potentially skewing the benchmark. Also, the B2000 proportions from each side of the corpus can be compared to assess the similarity of the somewhat small sample.

Here is a breakdown of the subject area coverage present in these 30 essays:

Table 2. *Texts comprising benchmark corpus*

Michigan 15 Essay Types	**BAWE 15 Essay Types**
Sociology 4	Sociology 3
History 2	History 3
Politics 2	Politics 3
Literature 2	Literature 1
Classics 1	Classics 1
Linguistics 1	Economics 1
Philosophy 1	Philosophy 1
Natural resources 1	Psychology 1
Biology 1	Anthropology 1
Total running words (tokens) = 29975	

None of the essays was viewed prior to selection, to limit bias. The mode of selection varied for each section of the benchmark corpus. The interface of the Michigan corpus is a website which allows user searches. The Michigan corpus search revealed 20 argumentative essays written by graduate native-speakers. Five of these were discarded due to brevity or large sections of archaic dialects and foreign language material. Then, an Excel spreadsheet of the BAWE corpus was searched by manually identifying essays written by first year undergraduate native-speakers and choosing them randomly from a list to generate a similar subject-area coverage to the essays from the Michigan part of the corpus. Together, these 30 essays represent a corpus of accomplished students' writing.

Lexical frequency profiles focusing on word *type* usage are sensitive to text length because, in longer texts, the proportion of words used only once will be much smaller relative to the most frequent words in the language. To overcome this, each text in the benchmark corpus was standardised to 1000 words by deleting all but the first 1000 words in each text. Cutting the texts in this way is perhaps not ideal as it assumes that the lexical profile of the first 1000 words will be consistent throughout the rest of the text, regardless of overall text length. The introduction, for example, a highly generic part of an academic essay may exhibit different proportions of frequent and infrequent lexis relative to the rest of the text. To my knowledge, however, this kind of variation has not been established empirically and using texts of varying lengths is not an option if word type proportions are to be measured.

Next, each of the 30 papers was analysed with AntWordProfiler (Anthony, 2014) using the NGSL and NAWL wordlists. Percentages are generated for each of the word lists. In addition, lexical items not included within NGSL and NAWL word lists ('off-list' words) were identified by AntWordProfiler and checked manually. The manual counts are necessary for off-list words in order to eliminate proper nouns from the count. An additional word list was used to exclude numbers from the main word lists. Then, F2000 proportion was calculated by adding together the 1k and 2k word types, and the B2000 proportion was calculated by adding all the remaining word types together, including the off-list types.

Proper nouns in this study were not reclassified as belonging to the most common one thousand words, as was done by Laufer and Nation (1995) but, instead, they were added to the

supplementary word count; that is, the words that do not fall into any category and are, therefore, not considered to be evidence of lexis at any level for the purposes of this study. Though neither method is without issue, adding proper nouns to the 1k word count would seem to inflate this level of lexical usage unnecessarily.

The participants and their academic context

In addition, to the benchmark corpus, essays were collected from the following two groups of participants:

IELTS 7.0 Group: 15 students on the International Foundation Programme (IFP) at King's College London between September 2016 and June 2017. The students are nationals from the following countries: Turkey (4), Kuwait (2), Egypt (2), Philippines (1), Mexico (1), Pakistan (1), UAE (1), Jordan (1) and Brazil (1) and Saudi Arabia (1). All of these students entered the IFP with an IELTS writing score of 7. These are among the highest writing scores of any students on our IFP. There were 21 students with this IELTS score in writing. In the end, only 15 of these 21 were available to participate so all those 15 were included in this group. Their overall IELTS score group mean is 7.6 and mean age for the group was 17.9 on entry to the foundation programme.

IELTS 5.5 Group: 15 students who entered the same course as above with IELTS writing scores of 5.5. This group is composed of nationals from China (6), Saudi Arabia (2), Turkey (2), Jordan (1), Uzbekistan (1), Japan (1), Pakistan (1) and Georgia (1). There were 63 students with this score in IELTS writing. The 63 potential participants were reduced to 15 by choosing 1 in 4 at random from a list of these students in alphabetical order until the 15 students had been selected. If the selected student was unavailable to participate or dropped out, I approached tutors of lower level classes to help me find available students with the appropriate score and contacted a limited number of students known to me. The average age of this group was 18.3 on entry to the foundation programme and their mean overall IELTS score was 6.1.

The foundation course contains a blend of tuition on study skills, academic language and subject-specific content. Dedicated vocabulary teaching occurs as a feature of three hour-long classes in terms 1 and 2 but also occurs at the teacher's discretion in other classes. Students are encouraged to learn vocabulary independently through their reading and exposure to

lectures. The same level of vocabulary input is timetabled for the higher and the lower level classes, although teachers may focus less or more on language when adapting the material to the specific needs of the group. Most, if not all, of the texts used on the IFP are professional-level authentic texts including published academic research newspaper articles. Texts are not adapted to have the overall burden of rare and difficult lexis decreased, whether given to higher or lower level students. Students vary considerably in their capacity for and engagement with independent learning. Additional studies with larger sample sizes might be able to reduce the potential influence of these factors on the results.

Samples were taken from the students' work in the same manner as for the benchmark corpus, with regard to standardising text length and calculating proportions of F2000 and B2000 words using AntWordProfiler. In addition, spelling mistakes were corrected but word choice errors of a morphological nature were left unchanged if the meaning was clear in the context. These were left, in part, due to the ambiguity of whether the error should be classified either as lexical or syntactic, as in the following example from essay from the IELTS 5.5 group:

> 'This phenomenon results in certain parts around the world **encountered** unparalleled growth and development in living standards'

Errors of this sort, however, were quite rare in the student's writing: less than one per paper on average. When words were used which did not make sense in the context, they were treated as unclassifiable, along with proper nouns.

This phase of data collection occurred in line with the assessment on the year-long programme. The first essay was submitted in week 9 of the course (Nov. 2016) and the second essay was submitted 18 weeks later (Mar. 2017). Both essays had an argumentative focus, in line with the texts in the benchmark corpus. The first essay was from a formative assessment in the students' optional module class. This sample of 30 1000-word essays comprises 18 Business Management, 4 Law, 4 International Relations and 4 Liberal Arts essays.

The second essay was submitted as a summative assignment in the Culture, Theory and Society module. While there were 5 different essay questions represented in this second sample, each question had an argumentative focus. Although it may have been preferable to

acquire data from texts in the same subject area for each essay, the structure of the course prevented that and it should be noted that the LFP measure has been found to be stable across different topics (Laufer & Nation, 1995).

Results

Data from the benchmark corpus

This benchmark consists of word *type* frequencies generated by AntWordProfiler. In the following table, the percentage of word types appearing at each level of the NGSL and NAWL are listed as mean scores for each 15-essay section of the benchmark corpus. The means represent average word-types in a 1000-word essay. They do not result from treating all fifteen essays as a single longer text, which would produce different results.

Table 3. *Mean word type percentages in benchmark corpus essays, by word list*

	NGSL1	NGSL2	NGSL3	NAWL	OFF LIST
Michigan 15	58.3	11.6	5.24	5.28	14.89
BAWE 15	56.43	13.05	5.87	4.83	13.56

The next step was to recast these figures as F2000 and B2000 proportions:

Table 4. *Benchmark corpus B2000 and F2000 proportions*

	MEAN F2000 Type %	MEAN B2000 Type %
Michigan 15	69.88 (5.63 SD)	25.42 (4.40 SD)
BAWE 15	69.48 (4.25 SD)	24.26 (5.12 SD)
Combined benchmark	69.68 (4.91 SD)	24.84 (4.73 SD)

The similarity of figures on each side the benchmark corpus suggests that similar figures might be found among similar samples, although this remains unsubstantiated at present. In addition, this 30-text sample passes the Shapiro-Wilk test (1965) for normal distribution and there is a 90% likelihood that another 30-essay sample would have a mean of between 23.44% and 26.24% (90% CI 23.44 to 26.24). The proportion of word types located within the 1k and 2k NGSL wordlists (69.68%) represents broader coverage than that achieved by the 1k and 2k original GSL wordlists (66.35%). This is what you would hope for from a more modern and

larger corpus. Interestingly, processing the same texts with the BNC/COCA 1k and 2k wordlists provides a similar but slightly lower figure: 69.14%.

The mean B2000 word type proportion from the benchmark corpus is 24.84% but there is a range of B2000 scores among the good native-speaker essays and the benchmark stated here aims to reflect that. Therefore, some lower values have been included from the distribution of B2000 scores to comprise the benchmark. These lower scores may represent a more achievable target whilst still representative of a B2000 proportion sufficient to 'have a positive effect on the reader' (Nation and Laufer, 1995:307) since *all* the essays in the corpus were graded with a merit, a distinction or an A grade. Although the institutions from which the essays come are unlikely to have standardized their grading procedures with one another, mean B2000 proportions are very similar, which may indicate a similar interpretation of the lexical features exhibited in these texts, to the extent that proportions of less common lexis contribute to higher grades.

The mean minus one standard deviation (4.73) was used as the lower end of the range of the benchmark and the mean was the higher end. Almost 90% of the 30-essay corpus (26 out of 30) had B2000 scores within or above this range.

Table 5: *Native-speaker benchmark as range between the group B2000 mean and -1SD*

Measure	**Mean**	**-1SD**
B2000 Word types	24.84	20.11

This figure has been slightly adjusted by the researcher for ease of use, rounding the entry score down to 20% and rounding up the mean to 25%. This range then, 20-25%, or beyond, is used in this study as the native-speaker B2000 benchmark. Non-native speaker texts will be considered to evidence a lexical level commensurate with successful native-speaker texts if the B2000 proportion falls within this range (or is higher).

Data from the students' essays

This table shows data from the IELTS 5.5 group's first and second essay. Scores for each essay are given as F2000 and B2000 proportions. B2000 proportions falling within the benchmark are highlighted:

Table 6. *IELTS 5.5 group F2000 and B2000 proportions in Essays 1 and 2*

	Essay 1		**Essay 2**		**B2000% change in essay 2**
Student	F2000%	B2000%	F2000%	B2000%	
1	81.44	13.85	79.24	15.45	1.6
2	79.8	15.02	72.69	17.56	2.54
3	71.94	15.59	66.5	27.25	11.66
4	75.53	17.49	70.34	17.94	0.45
5	76.64	17.3	76.21	15.85	-1.45
6	77.31	16.4	79.06	13.65	-2.75
7	83.66	13.37	75.24	16.67	3.3
8	75.89	20.82	75.57	20.6	-0.22
9	83.63	13.4	79.4	18.06	4.66
10	81.24	13.11	77.15	13.99	0.88
11	76.2	18.99	76.37	17.93	-1.06
12	71.18	15.85	74.5	20.4	4.55
13	75.66	18.55	82.01	15.87	-2.68
14	75.85	18.05	75.88	16.26	-1.79
15	75.62	17.73	76.27	19.61	1.88
Mean	**77.44**	**16.37**	**75.72**	**17.81**	1.44

In the earlier essays, only one had a B2000 proportion over 20%. In the later essays, that figure had risen to three. Of the 15 later essays, six did not contain an increased proportion of B2000 word types and the overall group mean increase in B2000 words, 1.44%, seems quite small. A paired-samples two-tailed T-test on the B2000 proportions in essays 1 and 2 returned a non-statistically significant result ($p=0.16$) and the effect size is small ($r^2=0.14$).

In fact, the IELTS 7.0 group evidence much greater apparent lexical development:

Table 7. IELTS 7.0 group F2000 and B2000 proportions in Essays 1 and 2

	Essay 1		Essay 2		B2000% change in essay 2
Student	F2000%	B2000%	F2000%	B2000%	
1	75.27	19.78	71.96	19.26	-0.52
2	73.82	20.99	73.17	21.33	0.34
3	78.14	14.82	69.33	21.45	6.63
4	83.01	15.06	61.11	30.56	15.5
5	80.85	16.17	69.32	20.31	4.14
6	78.53	18.59	71.02	18.8	0.21
7	81.18	13.44	75.36	16.83	3.39
8	79.3	16.48	68.53	25.13	8.65
9	79.08	14.46	78.38	18.31	3.85
10	73.59	20.29	74.15	20.74	0.45
11	75.35	18.97	71.7	21.82	2.85
12	65.1	24.38	70.21	27.28	2.9
13	73.63	18.91	75.36	20.05	1.14
14	70.56	25.46	63.98	31.98	6.52
15	71.1	21.61	71	23.75	2.14
Mean	**75.9**	**18.63**	**70.98**	**22.51**	3.88

The two groups were not that far apart in terms of the mean B2000 proportions evident in the first essay (16.37% and 18.63% respectively). Given that Cambridge English (2018) consider a 1.5 band difference in overall IELTS score (which is the difference between these groups) equivalent to a whole CEFR band difference, this 2.26% difference is smaller than expected. However, a big difference was evident when it came to how much increase in B2000 word types was apparent in the second essay: a 3.88% increase for the higher-level group. The T-test using this group's B2000 scores from essays 1 and 2 (paired-samples; two-tailed) returned a statistically significant result ($p=<0.05$) and a large effect size ($r^2=0.48$), suggesting that the linguistic intervention represented by the 18-weeks of tuition between essays 1 and 2 had been more effective for this higher group.

Moreover, the group mean of 22.5% places the IELTS 7.0 group mean firmly within the native-speaker benchmark, as can be seen in Figure 1. 11 out of 15 texts fall within the benchmark and, of the remaining 4, 3 are less than 2% outside it. In short, almost the whole IELTS 7.0 group is operating within or near the benchmark at the end of the foundation year.

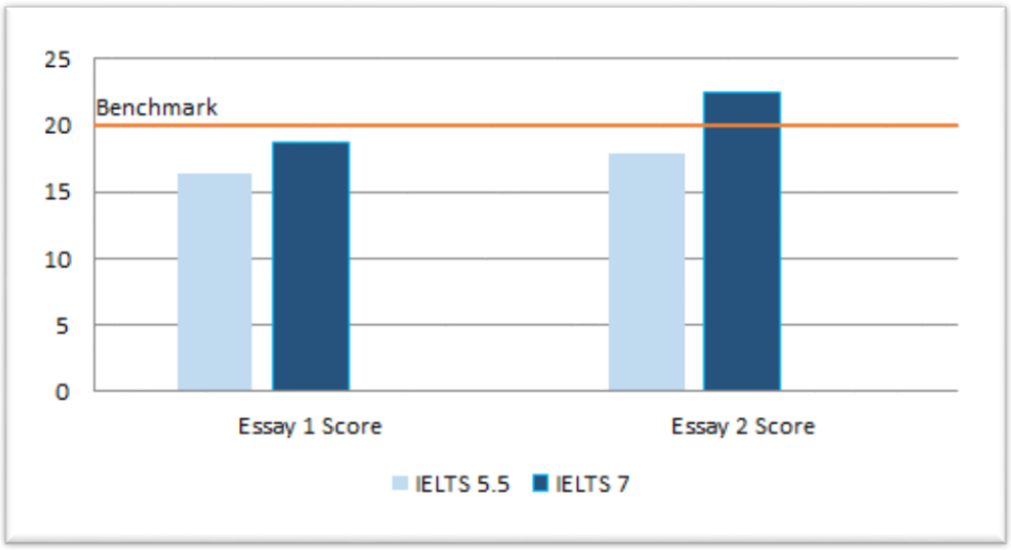

Figure 1. *Mean B2000% for first and second essay*

The lesser increase for the IELTS 5.5 group is a somewhat disappointing amount of progress; it might have been hoped that the relatively small gap in B2000 productive lexis in the first piece of writing (2.26%) would not have increased to the larger gap evident in the second essay (4.7%). The B2000 proportion, however, is only one measure of lexical level and, as outlined in the literature review, there are others. The apparent lack of progress highlighted by the B2000 measurement led the researcher to reassess the IELTS 5.5 texts using TTR to see if other kinds of progress existed. This was calculated by dividing the total number of word types in each 1000-word essay, by the total number of running words (as calculated by AntWordProfiler), once the proper nouns had been removed from the word type count. Indeed, applying the TTR measure to the same texts suggests a slightly enhanced picture of the lexical progress for some students:

Table 8. *TTR from essays 1 and 2 for IELTS 5.5 group*

Student	Essay 1	Essay 2	TTR change in second essay
1	0.35	0.38	0.03
2	0.38	0.38	0
3	0.38	0.38	0
4	0.41	0.39	-0.02
5	0.41	0.38	-0.03
6	0.38	0.40	0.02
7	0.39	0.37	-0.02
8	0.36	0.39	0.03
9	0.4	0.42	0.02
10	0.37	0.41	0.04
11	0.33	0.36	0.03
12	0.34	0.44	0.1
13	0.39	0.37	-0.02
14	0.39	0.36	-0.03
15	0.34	0.40	0.06
Mean	**0.375**	**0.389**	**0.014**

The mean TTR for the IELTS 5.5 group was 0.375 for their first essay and 0.389 for their second essay. In a paired-samples T-test (two-tailed), the difference between the essay 1 and essay 2 TTR scores is not statistically significant ($p=0.28$), with a very low effect size ($r^2=0.08$). However, three out of the six writers who had produced a second essay with a *decrease* in B2000 proportion show an *increase* in TTR (highlighted in table 8). The increase noted here is relatively small: in a 1000-word essay, a change of 0.014 indicates a change an additional 14 unique words. But for there to be an increase in TTR across two essays, there needs to be an increase in the overall number of word types used. If that increase has not occurred in the B2000 proportion, it will have occurred at the 1k and 2k level. For some students in this group, productively using a wider range of word types from the 1k and 2k level may represent the most appropriate next step in acquiring a productive vocabulary. However, the increases noted by these two measurements made the lower level group (B2000 increases of 8.8% and TTR increase of 3.73%) would seem to equate to less overall progress than the higher group (B2000

increase of 20.83%). If the foundation programme were considered a linguistic intervention, it was a more effective one for those who entered the programme with higher-level skills.

Discussion

Both groups of students in this study displayed an increase in the proportion of B2000 word types in their writing, similar to the findings of Turlik (2013) and Mazgutova & Kormos (2015). However, the disparity in the overall level of progress and the type of progress evident between the two groups has implications for the structuring of foundation programmes. Whilst lexical instruction on foundation programmes may begin with the assumption that F2000 word families need *not* be an explicit focus for any group on the programme, data from this study suggest otherwise. Assuming basic vocabulary is known and using the AWL, or other such B2000 word list, as a starting point for adding to IELTS 5.5 students' productive lexical knowledge may not be appropriate for some students within this band. The evidence for this is the students who did not increase their B2000 proportion but did increase the overall number of word types in their writing. An assessment of the students' productive lexical abilities at the 2k level could help here. If there is flexibility in structuring the course to allow for additional lexical input for the lower level groups, rather than presenting lower and higher ability groups with the same curriculum, then that would be advisable. The content the IELTS 5.5 group were required to process in this foundation year was mostly authentic academic texts (as stated above), not edited for the level of the student and it ought to be investigated, if ethical, how detrimental a heavy burden of unknown lexis is to lexical acquisition, relative to a more manageable load.

The robustness of an IELTS 7.0 writing score as an indicator of linguistic readiness for academic study is questioned by the results of this study, in line with the findings of Drummond (2018) and Cooper (2013). Essays receiving merit grades and above typically have at least 20% B2000 word types, according to the benchmark data from this study. The IELTS 7.0 group in this study would have struggled to produce B2000 vocabulary in this proportion, had they gone directly into undergraduate study, as is evident in their first piece of writing. Whilst many IELTS 7.0 students may cope on direct entry to undergraduate or post-graduate programmes, if there is an opportunity to intentionally enhance their productive lexical abilities, these results argue some students at this level would benefit from it. Drummond (2018) also shows a wide range of receptive lexical knowledge within the IELTS 7.0 overall band. Taken together, the studies present a picture of considerable variation of lexical competence within the band and argue

against the assumption that an IFP curriculum can ignore lexical development for this type of student.

As described earlier, the LFP is, in some ways, a more nuanced measure of productive lexical ability than TTR, and TTR has been become an unfashionable measure. However, as is evident in this study, the B2000 measure, if used alone, may obscure a certain kind of progress in productive lexical ability. The problem is essentially this: the B2000 measure does not give credit if *the total number of unique words* used in a piece of writing has increased but *the proportion in each frequency band* has not changed. Consider the following data on student 6 from the IELTS 5.5 group:

Table 9. *A decrease in B2000 with an increase in overall word types*

Essay	**B2000%**	**Total no. of word types**
1	16.4	377
2	13.65	397

Her second essay, if only assessed in terms of its lower B2000%, would not show that additional productive capabilities are emerging at the 1k level and the 2k level, and that the total number of unique words has increased by 20 in total. Looking at data relating to whole batch of essay from the IELTS 5.5 group illustrates a similar point:

Table 10*. A decrease in 1k and 2k proportions, with an increase in overall word types*

Essay	**1k proportion**	**1k word types**	**2k proportion**	**2k word types**
1	45.43	1159	18.54	473
2	42.08	1191	17.67	500

The 1k and 2k proportions do not decrease much in the second essay, but using the B2000 measure alone would not give credit for an increase in the total number of word types used in these higher frequency bands. The partially discredited measure of TTR would be able to broadly indicate progress of this sort. In this study TTR proved useful in terms of establishing progress for some students not evident with the LFP/B2000 ‘lens’. TTR, then, should perhaps not be discarded by researchers in lexical development but retained as a means of corroborating other measures.

A tool for teachers

The process of comparing EAP students' writing to the 20%+ B2000 benchmark can be done with lextutor.ca (Cobb, 2016) and the benchmark data and simplified methods from this study. EAP practitioners may wish to do this as a means of initial or diagnostic assessment. It is more straightforward to use the F2000 data as a benchmark for this purpose as this figure does not need to be recalculated after the removal of proper nouns. The native-speaker F2000 mean from 30 essays is 69.7% with an SD of 4.91. Rounding these figures a little creates the range of 70%-75% word types in native speakers' essays from 1k and 2k word frequency levels. For assessment purposes, any F2000 score less than 75% is within the range of native-speaker lexical usage evident in highly-graded essays and as this figure decreases, so the proportion of less frequent, potentially academic or specialized vocabulary increases. Practitioners wishing to do this need to standardize the essay length to 1000 words and select the NGSL/NAWL word lists within the *vocabprofile* program on lextutor.ca. Data on word *types* rather word *tokens* needs to be checked and cumulative scores compared with the benchmark stated in this study.

Limitations of the study

The LFP measure regards lexis as discrete, individual items. Whilst there is certainly merit in this, developing lexical proficiency is also a matter of collocation, lexical bundles, and a range of other types of fixed and semi-fixed expressions. Future research could produce a more accurate picture of the lexical level and lexical development in L2 student academic writing if it were able to create a synthesis of these elements. The methodology used does not discriminate between items which have been used effectively and those which have been used in a non-standard context or with non-standard collocations. Below is an extract from a text from IELTS 5.5 group which exhibited a high B2000 word type proportion but was, at times, difficult to understand. The B2000 tokens are in bold:

> '**Proponents** of **globalisation** observed **globalisation** as being **evolutionary** and **assisting**, while critics consider **globalisation** as **colonising** and the cause of **relapsing** our modern world.'

Here, the constituent words are relatively infrequent, but the overall message is difficult to understand due to non-standard collocations around *evolutionary* and *relapse*. From reading the

essays in this study it seems probable that a measurement which considered the appropriacy of usage across the two groups' writing would show that the already significant gap in lexical knowledge was in truth greater still.

Conclusion

Using IELTS writing scores at course entry as the organising principle, two groups of participants were studied: an IELTS 5.5 group and an IELTS 7.0 group. Two essays from each group were processed by AntWordProfiler to establish the mean B2000 word type proportion of each group near the beginning and end of a one-year international foundation programme. Comparing this data with the 20%+ B2000 native-speaker benchmark yields the following results: the IELTS 7.0 group began with mean B2000 word type proportion lower than the benchmark but the group mean is firmly within this benchmark by the end of the year. The IELTS 5.5 group began relatively close to the IELTS 7.0 group but did not add B2000 vocabulary to their essays at the same rate as the higher group. Some additional lexical development occurred at the F2000 level for the IELTS 5.5 group.

For the IELTS 7.0 group, the foundation year was required before their B2000 word type proportion resembled the native-speaker benchmark. The IELTS 7.0 group apparently developed B2000 lexis more rapidly than the IELTS 5.5 group, perhaps indicating that the material presented on the course was more conducive to facilitating additional lexical production for the higher group. Lexical development seemed to occur at the F2000 level for some within the IELTS 5.5 group, suggesting that some F2000 lexis could precede the introduction of academic word lists on similar foundation courses. The increase in the overall number of word types used by some students in the lower level group would not have been noted by using the B2000 measure exclusively, indicating that other measures such as TTR could be used alongside it to identify increases in the overall number of unique words used in a text.

End Note

Warwick University request that the following text accompanies any study utilising the BAWE corpus: 'The data in this study come from the British Academic Written English (BAWE) corpus, which was developed at the Universities of Warwick, Reading and Oxford Brookes under the directorship of Hilary Nesi and Sheena Gardner (formerly of the Centre for Applied Linguistics, Warwick), Paul Thompson (formerly of the Department of Applied Linguistics, Reading) and

Paul Wickens (School of Education, Oxford Brookes), with funding from the ESRC (RES-000-23-0800).'

Biodata

Andrew Drummond is an English for Academic Purposes teacher at King's College London. He has previously taught in South Africa, Hungary and Macedonia (FYROM). His research interests include academic vocabulary, the link between vocabulary knowledge and assessment, and the link between vocabulary knowledge and reading skills. Andrew can be contacted at andrew.drummond@kcl.ac.uk and @drummondandrew on Twitter.

References

Anthony, L. (2014). *AntWordProfiler* (Version 1.4.1) [Computer Software]. Tokyo, Japan: Waseda University. Accessed 15/07/2016. Available from http://www.laurenceanthony.net/

Bauer, L. & Nation, P. (1993). Word families. *International Journal of Lexicography*, *6*(4), 253-279.

BAWE—Available at http://www2.warwick.ac.uk/ fac/soc/al/research/collect/bawe/. Accessed 9/7/2016.

Brown, D. (2012). Exploring L2 learners' productive knowledge of collocations. *45th Annual Meeting of the British Association for Applied Linguistics*. Accessed 20/08/2016 Available from: https://www.academia.edu/1961845/Exploring_L2_learners_productive_knowledge_of_collocations

Browne, C. (2014). A New General Service List: The Better Mousetrap We've Been Looking for?. *Vocabulary Learning and Instruction*, 3(1), 1-10.

Browne, C., Culligan, B. & Phillips, J. (2013). The New General Service List. Accessed: 16/05/16. Available from http://www.newgeneralservicelist.org.

Cambridge English. (2018). *Comparing scores to IELTS.* Accessed 18/03/10. Available from: https://www.cambridgeenglish.org/Images/461626-cambridge-english-qualifications-comparing-scores-to-ielts.pdf

Cobb, T. *Web Vocabprofile*. Accessed 01/7/2016. Available from http://www.lextutor.ca/vp/

Cooper, T. (2013). Can IELTS writing scores predict university performance? Comparing the use of lexical bundles in IELTS writing tests and first-year academic writing. *Stellenbosch Papers in Linguistics Plus*, 42, 63-79.

Corson, D. (1997). The learning and use of academic English words. *Language Learning*, *47*(4), 671-718.

Coxhead, A. (2000). A new academic word list. *TESOL Quarterly*, 34(2), pp.213-238.

Crossley, S. A., Salsbury, T. & Mcnamara, D. S. (2015). Assessing lexical proficiency using analytic ratings: A case for collocation accuracy. *Applied Linguistics*, 36 (5): 570-590

Dooey, P. & Oliver, R. (2002). An investigation into the predictive validity of the IELTS Test as an indicator of future academic success. *Prospect,* 17, 36 -54.

Drummond, A. (2018). Investigating the Relationship between IELTS Scores and Receptive Vocabulary Size. *Journal of the Foundation Year Network*, 1.

IELTS. (2015). *Guide for teachers.* Accessed 02/05/2017. Available from: https://www.ielts.org/-/media/publications/guide-for-teachers/ielts-guide-for-teachers-2015-uk.ashx

Gregori-Signes, C. & Clavel-Arroitia, B. (2015). Analysing Lexical Density and Lexical Diversity in University Students' Written Discourse. *Procedia-Social and Behavioral Sciences*, 198, 546-556.

Halliday, M., Matthiessen, C. M. & Matthiessen, C. (1985). *An introduction to functional grammar.* London: Routledge.

Heatley A., Nation ISP & Coxhead A. (2002). Range [Computer software]. Accessed 05/06/2017. Available from http://www.victoria.ac.nz/lals/staff/paul-nation/nation.aspx.

Hyland, K. & Tse, P. (2007). Is there an "academic vocabulary"? *TESOL Quarterly,* 41(2), 235-254.

Iwashita, N. (2005). An investigation of lexical profiles in performance on EAP speaking tasks. *Spaan Fellow Working Papers in Second or Foreign Language Assessment, 3,* 101-111

Laufer, B. (1995). Beyond 2000: a measure of productive lexicon in a second language. In: Eubank, L., Selinker, L., Sharwood Smith, M. (Eds.), *The Current State of Interlanguage*. Amsterdam: John Benjamins BV, 265–272.

Laufer, B. & Nation, P. (1995). Vocabulary size and use: Lexical richness in L2 written production. *Applied Linguistics*, 16(3), 307-322.

Lemmouh, Z. (2008). The relationship between grades and the lexical richness of student essays. *Nordic Journal of English Studies*, 7(3), 163-180.

Loewen, S. & Plonsky, L. (2015). *An A–Z of Applied Linguistics Research Methods*. New York: Palgrave Macmillan

Mazgutova, D. & Kormos, J. (2015). Syntactic and lexical development in an intensive English for Academic Purposes programme. *Journal of Second Language Writing*, 29, 3-15.

Römer, U. & Swales, J. M. (2010). The Michigan corpus of upper-level student papers (MICUSP). *Journal of English for Academic Purposes*, *9*(3), 249.

Schmitt, N. (2010). *Researching vocabulary*. Basingstoke, England: Palgrave Macmillan.

Schmitt, N. & Zimmerman, C. (2002). Derivative word forms: what do learners know? *TESOL Quarterly, 36*(2), 145-171.

Shapiro, S. S. & Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, *52*(3/4), 591-611.

Singh, P. & Doherty, C. (2004). Global cultural flows and pedagogic dilemmas: Teaching in the global university contact zone. *TESOL Quarterly*, *38*(1), 9-42

Stoeckela, T. & Bennett, P. (2015). A Test of the New General Service List. *Vocabulary Learning and Instruction*, *4*(1), 83-94

Treffers-Daller, J., Parslow, P. & Williams, S. (2016). Back to basics: how measures of lexical diversity can help discriminate between CEFR levels. *Applied Linguistics*, doi: 10.1093/applin/amw009

Turlik, J. (2013). The use and development of academic vocabulary in L2 writing: a longitudinal investigation. *Learning and Teaching in Higher Education: Gulf Perspectives*, *10*(1) 1-14

West, M. (1953). *A general service list of English words*. London: Longman, Green.

Xue, G. & Nation, I. S. P. (1984). A university word list. *Language Learning and Communication*, *3*(2), 215-229.

Yu, G. (2010). Lexical diversity in writing and speaking task performances. *Applied Linguistics*, *31*(2), 236-259.